

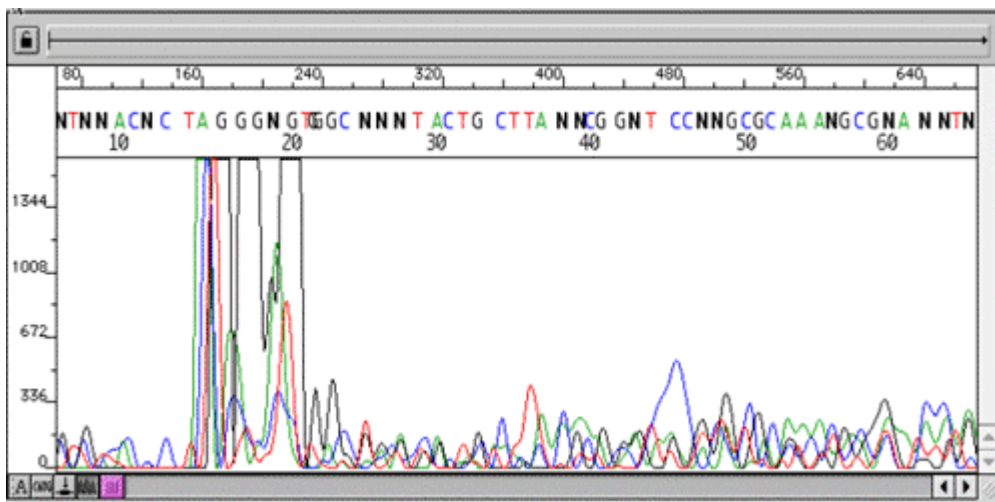


# Troubleshooting Your Data

This document was taken on the Roswell Park Institute website  
<http://www.roswellpark.org>

The two most common causes for failure to get good or any sequence data for your samples are purity and concentration of your template DNA. If it appears that you have done everything correctly, then look below for some additional reasons why you might obtain less than optimal DNA sequence data quality. We've listed various causes, solutions and, for some, pictorial representations of what these specific problems might look like. Many causes and solutions may look rather obvious and just involve common sense, but you'd be surprised how many times we've heard "how could I have done that?"...

## 1. No sequence data



**Cause:** not enough or no DNA/primer in tube

**Solutions:** doublecheck your quantitations, stock concentrations and dilutions.. While our sequencers are very sensitive and can detect a range of DNA concentrations, there is still a "threshold" amount that must be reached to obtain any sequence data.

**Cause:** inhibitory contaminant

**Solutions:** the cycle sequencing reaction used to amplify samples for automated sequencing is very sensitive to the presence of certain contaminants, some of which will completely inhibit our sequencing enzyme. Please check the Contaminant section for a list of potential inhibitors. You may need to reprep your sample to sufficiently remove one or more inhibitory components to obtain any sequence data.

**Cause:** priming site not present

**Solutions:** if you've chosen one of the sequencing facility's vector primers (T7+, T7-, T3, SP6,

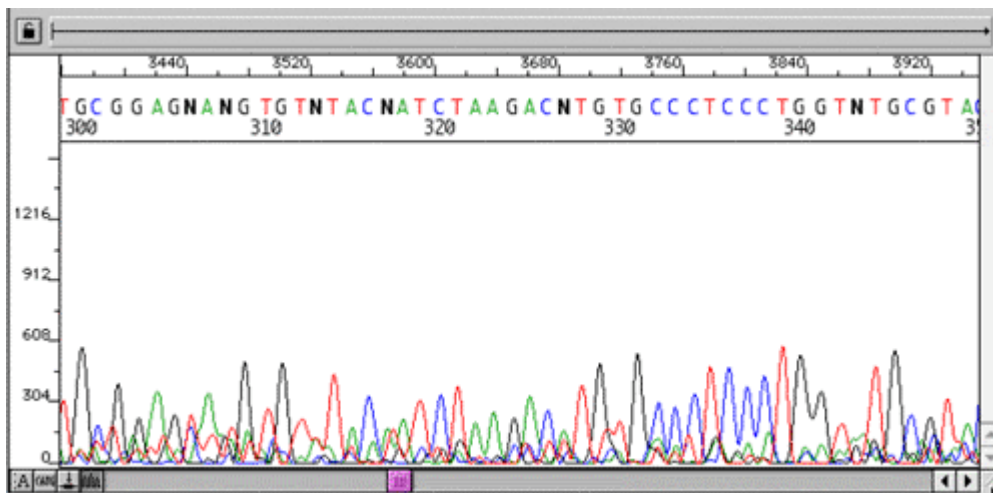
M13+, M13-), make sure it is present in your vector. While many of the primers we provide are quite common to many different vectors. Doublecheck your plasmid maps/sequences.

- if you've designed your own custom primer from previous sequence data, make sure you were using a reliable area of sequence - look for sharp, well-defined peaks with no ambiguity. Avoid areas where the peaks are broader and not well separated - this will occur towards the end of the sequence where the fragments are larger and the polymer cannot adequately resolve single nucleotides, causing inaccurate basecalling.

**Cause:** expired reagents

**Solutions:** falls under common sense category

## 2. Noisy data with weak signal



"Noisy" data can be identified by the presence of multiple peaks and numerous "N"s within your sequence. The Sequencing Analysis program assigns an "N" as a base identification when there are two or more peaks present at one position. This "N" may signify the legitimate occurrence of two nucleotides, as in the case of a heterozygote, but may also be seen when background noise is high or when multiple products are present. When your sample exhibits weak signal, the software attempts to compensate by boosting up the signal of sample bands to detectable levels. However, the background noise will also be artificially amplified, giving a poor signal-to-noise ratio. Background noise appears as many smaller, undefined peaks under your sequence peaks of interest. This noise is always present, but with well-prepared samples of good signal strength, it will be undetectable. To determine if your noisy data may be due to weak signal, look at your ABI trace file. If you are looking at a paper chromatogram, look towards the top and middle of your trace for a line that says "Signal". If the file is on your computer, click the "A" radio button in the bottom left-hand corner, which is visible when you have opened up the trace file within a viewing program, such as EditView or Chromas. Scroll down to the line that says "Signal" and you will see the four nucleotides followed by numbers in parentheses. These numbers represent the average signal strength of each nucleotide and their values should, optimally, be between 200-400. If they are much less than 100, then you can assume your noisy data is at least partially due to its weak signal.

**Cause:** not enough DNA

**Solutions:** doublecheck your quantitations, stock concentrations, calculations and dilutions. Make sure you've provided the appropriate amount of DNA and/or primer.

**Cause:** inhibitory contaminant e.g..salts, phenol

**Solutions:** the cycle sequencing reaction used to amplify samples for automated sequencing is very sensitive to the presence of certain contaminants, some of which can partially or completely inhibit our sequencing enzyme. You may need to re-purify your sample to sufficiently remove one or more inhibitory components to obtain better sequence data.

**Cause:** degraded DNA from nucleases, repeated freeze-thaw, excessive UV light exposure, bisulfite treatment.

**Solutions:** Nuclease contamination in a template preparation as well as repeated freeze-thaw cycles can degrade DNA over time. Even low amounts of nucleases can extensively degrade DNA depending on storage conditions and temperatures, as well as the length of time the DNA is stored. Generally, re-isolation and purification of the template DNA will be necessary to obtain good DNA sequence. When extracting PCR products from a gel, prolonged exposure to UV light will degrade and nick the DNA. Limit the time and UV intensity as much as possible to prevent degradation. When treating DNA with bisulfite for methylation experiments, it is important to avoid long incubations at higher temperatures as substantial amounts of DNA will be degraded in this process.

**Cause:** trend in worsening data?

**Solutions:** if you have previously been able to obtain good sequence data but begin to see a deterioration in quality that gets progressively worse, you may have some contamination in one or more reagents, or have some reagents that have reached the end of their usefulness. Make up fresh stocks of commonly used reagents, such as buffers, and always use high quality distilled water in your preparations.

**Cause:** inefficient primer binding (low T<sub>m</sub>, degenerate primers, mismatch)

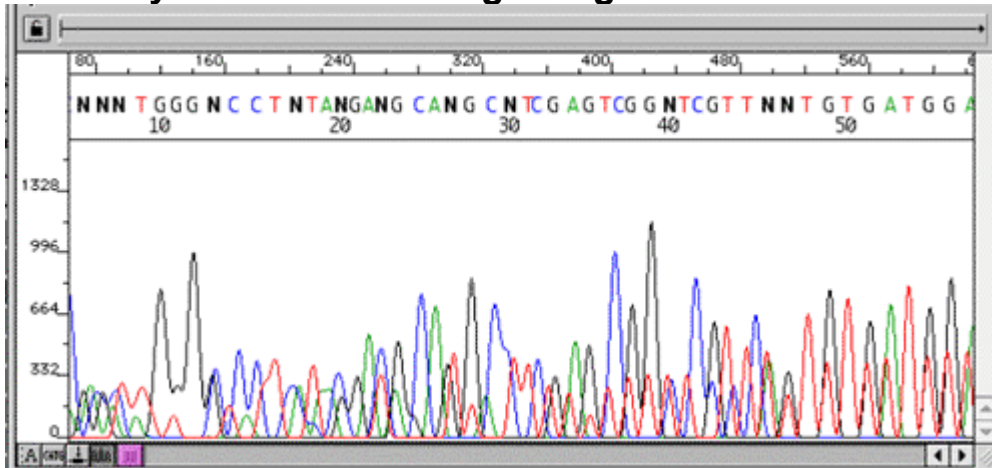
**Solutions:** the T<sub>m</sub> of a primer is defined as the temperature at which 50% of the oligonucleotide and its perfect complement are in duplex. The T<sub>m</sub> of an oligo can be roughly calculated by using the formula:

$$T_m = 2^{\circ}\text{C}(\text{A}+\text{T}) + 4^{\circ}\text{C}(\text{G}+\text{C})$$

This is the most commonly used formula for calculating T<sub>m</sub>, though it is not the most accurate as it does not factor in salt or formamide concentrations. A good website to check out if you are interested in some detailed theory behind T<sub>m</sub> calculations is [http://www.sigma-genosys.com/oligo\\_meltingtemp.asp](http://www.sigma-genosys.com/oligo_meltingtemp.asp).

In our cycle sequencing reaction, our primer/template annealing step occurs at 50°C. Thus, if your primer T<sub>m</sub> is much lower than 50°C, hybridization to its complementary template will be much less efficient and a lesser number of extending fragments will be generated. Increase your primer T<sub>m</sub> by adding additional bases to the 5' or 3' end to raise the T<sub>m</sub> to be within the range of 52°C-58°C. Degenerate primers and those with mismatched bases will also show decreased hybridization efficiency due to reduction of the stability of primer binding, and if degeneracy or mismatches occur at or near the 3' end of your primer, it is highly likely that your sequencing attempt will fail.

## 2.1 Noisy data From the beginning



**Cause:** multiple priming sites involving vectors

**Solution:** your primer may have a secondary hybridization site that may be identical or closely related, with different nucleotide sequences following each site, giving superimposed bands within your sequence. If the priming sites are identical, (such as when more than one T7 promoter site is present, for example), the double peaks will be strong from the outset. The fragments may also show shifted migration so that the double peaks are not directly on top of one another but will be offset to one side or the other due to the differing mobility patterns of the strands with dissimilar nucleotide composition. In other instances, a secondary priming site may not be exactly the same, but may differ by a few internal bases. In this case, the mismatched primer may not hybridize as efficiently but can still anneal and extend, and give rise to less intense fragments that can be seen underneath your peaks of interest. In both cases, it's necessary to screen both your vector and insert carefully to look for sequences that may match or be similar to your proposed primer. You may need to choose another vector primer on the same end of the multiple cloning site or redesign your custom primer. When choosing another primer is difficult, such as when primer walking through a repetitive area, try to find a primer that has a 3'-base match specific to your area of interest which can help act as an "anchor".

**Cause:** multiple priming sites in PCR

**Solution:** this may occur when one or both of the PCR primers hybridizes to more than one position on the template DNA, giving rise to multiple PCR products. Often this will be obvious when visualizing the PCR products on an agarose gel as there will be more than one band present. In this case, gel purification of the desired product will be necessary. One can run into difficulty, however, when the products are very similar in size, which may arise when amplifying related or repetitive DNA, and do not separate well on the gel. In this case, optimization of the PCR reaction may be necessary or redesign of the PCR primers in order to choose a more specific priming site.

**Cause:** PCR primers acting as both forward and reverse

**Solution:** sometimes, a PCR product may be generated when one primer functions as both the forward and reverse primer in the PCR reaction, giving rise to an artifactual product. This is fairly easy to detect when sequencing the PCR product as one primer will give double peaks from the start, while the other fails to give any sequence data. Redesign your set of PCR primers.

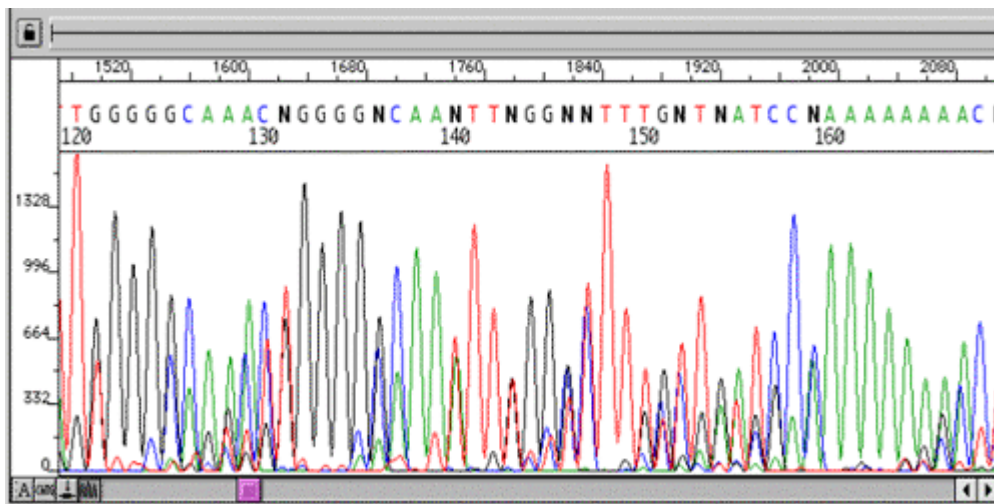
**Cause:** residual PCR primers and/or dNTPs

**Solution:** as two primers are present in the PCR reaction, incomplete removal of these primers can lead to double peaks within the sequencing data. Both primers will act as sequencing primers and lead to superimposed bands which correspond to the complementary strands from opposite orientations. It is critical to remove excess primers and dNTPs from the PCR reaction by purification. If attempting to do direct sequencing of PCR products without purification by diluting an aliquot of your PCR product with water to lower the concentration of residual primers and dNTPs (a method which we do not recommend), then it is imperative to optimize your PCR reaction so that primers and dNTPs are used in limiting amounts so that most are used up by the end of the PCR.

**Cause:** primers with high T<sub>m</sub>

**Solution:** primers that have a T<sub>m</sub> much higher (>65°C) than our suggested 52°C-58°C often do not function well as sequencing primers. When primers have a T<sub>m</sub> that high, it is often a result of increased G-C content or because the primer is quite long, both factors that can increase the potential for primer secondary structure formation. If possible, choose another primer with a lower T<sub>m</sub>.

**Cause:** primers with n-1 population

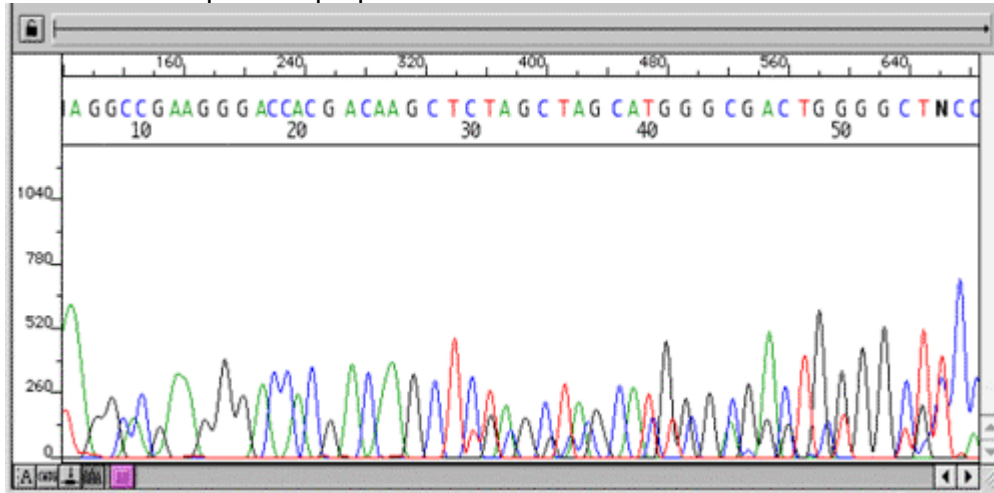


**Solution:** this problem is not uncommon and can result from poor quality synthesis of sequencing primers. Primers are synthesized from the 3'end to the 5'end and when synthesis is inefficient, there can be a significant population of less than full-length primers - n-1s, which are full-length primers minus one base, plus other shorter derivatives. These primers have a common 3'end but different 5'ends, thus chains that terminate at the same position will have different lengths and will run at different positions on the gel. Primers that have degraded from the 3'end will also give this appearance. It is easy to spot this problem within the sequencing chromatogram as each position will contain the true peak as well as the peak immediately to the right of it, giving the appearance of "shadow" peaks. Whatever the cause of the n-1s, it will be necessary to resynthesize the primer to obtain an oligo of suitable quality for sequencing. When high-quality reagents and proper protocols are utilized during oligo synthesis, cartridge or HPLC purification of the primers is usually not necessary for typical oligos (<30 bp), but sometimes additional purification can be beneficial.



## 2.2 Noisy data Begin farther into the sequence

**Cause:** mixed plasmid prep



**Solution:** a plasmid prep that is contaminated by more than one product, such as two vectors with different inserts or vector with insert and vector without, will generally show an early section of clean sequence data (common vector multiple cloning site sequence) followed by double peaks. Occasionally, a plasmid may contain more than one vector molecule or may encounter spontaneous deletions or insertions during growth. The point at which the double peaks begin corresponds to the start of the insert cloning site. To avoid this problem, it's important to carefully pick a single colony from your growth plate, restreaking if necessary, to be sure that your colony is completely clonal. You should follow this up with a restriction digest of your plasmid run out on an agarose gel to ensure vector and insert are present as expected.



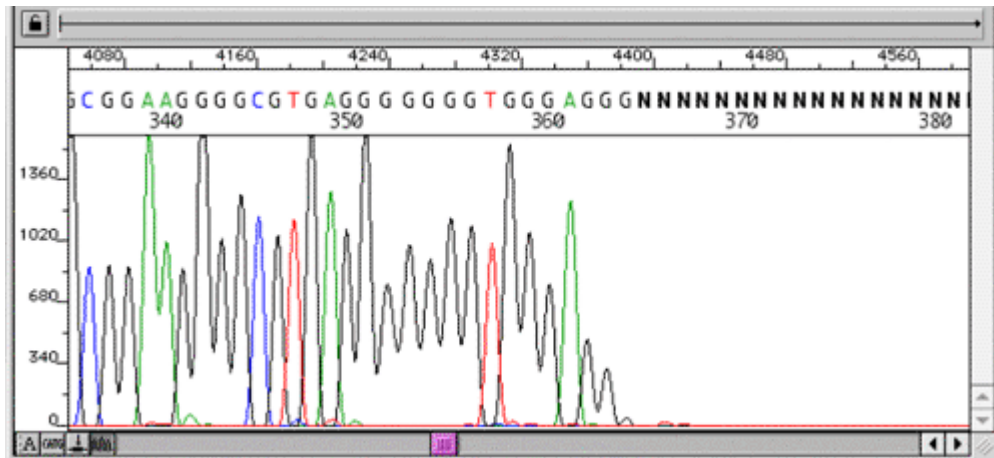


necessary to re-isolate your DNA to get a pure clone containing only one of the molecules. With PCR products, you will need to gel purify the two products in order to separate them.

## 4. Truncated sequences

Truncated sequences can be characterized as abrupt or gradual. Abrupt truncations will show strong, clean signal up to a point and then drop sharply down over the course of a few nucleotides to much weaker or no detectable signal. Gradual truncations will show good sequence data initially but then begins to taper off to progressively weaker, smaller peaks until there is nothing but background noise. The nature of the truncation can sometimes help to determine its cause.

**Cause:** secondary structure



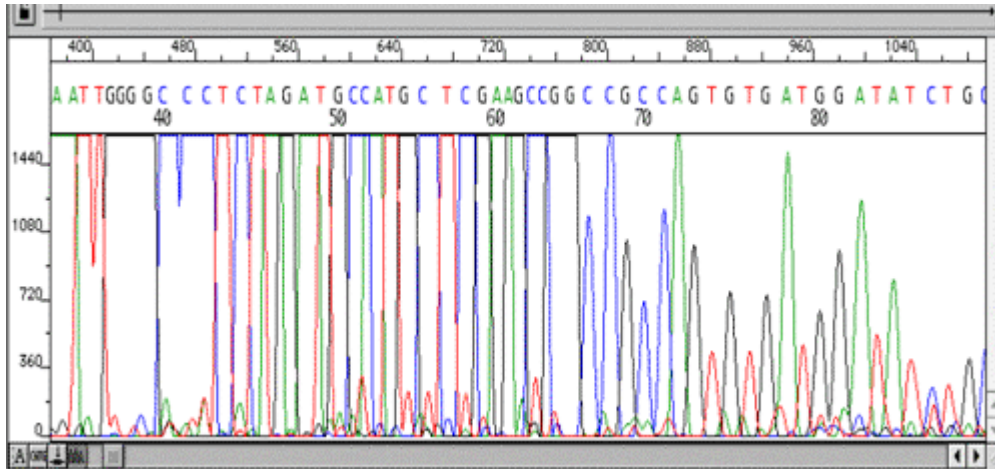
**Solution:** G-C rich, and to a lesser degree, A-T rich, DNA is predisposed to secondary structure formation, as strong hydrogen bonding between G and C nucleotides can cause the template DNA to loop or bend and anneal to complementary sequences, forming hairpins that can restrict the passage of the sequencing polymerase and thus be very difficult to sequence through reliably. These hairpins may not melt at our cycle sequencing temperatures and can cause premature termination of sequence data. Secondary structure may appear as a sharp termination of signal with no sequence data after, or if the loop has been relaxed slightly, you may see strong signal that drops abruptly but may have some weaker peaks following that are still quite accurate. With the newest formulation of BigDye Terminator chemistries (v3.1), some G-C rich difficulties have improved dramatically, but unfortunately it hasn't solved everything. There is not one solution that resolves every secondary structure problem, but there are couple you can try and usually one will allow us to read through it. Placing a primer as close to the hairpin loop as possible to help force its unwinding has also worked in the past. Sequencing the opposite strand can sometimes lead to a huge improvement. If these solutions don't work, we may suggest you try linearizing your DNA with restriction enzymes to help relax the hairpin. And if you are trying to PCR up a very G-C rich region, addition of betaine or DMSO to your PCR reaction can help, as can substitution of 7-deaza dGTP for 75% of the dGTP in your PCR reaction. And if all else fails, you can try manual radioactive sequencing as a last resort.

**Cause:** linearized DNA

**Solution:** if your DNA has been cut with one or more restriction enzymes, the sequence data will sharply end at the recognition site of the enzyme that cut at the 3' end of your insert. Did you

accidentally send us digested DNA? Run it out on a gel to see.

**Cause:** too much DNA

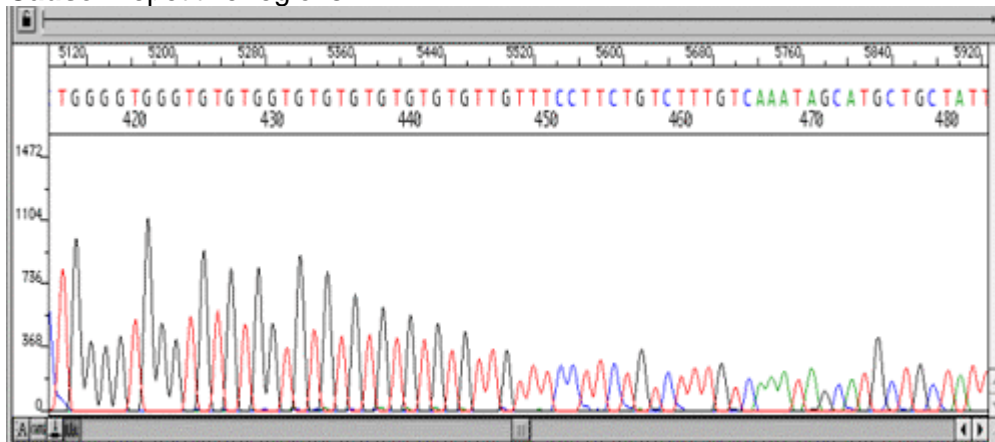


**Solution:** while there is a range of DNA concentrations we can sequence reliably, too much DNA will cause premature termination of signal. Overloading of DNA will exhibit early top-heavy peaks followed by rapidly weakening peak height and strength. This occurs because the dNTPS in the cycle sequencing reaction will be distributed among too many extending chains and will be depleted early on, resulting in an excessive amount of short fragments.. In addition, if your template is impure, higher concentrations of DNA can be accompanied by higher amounts of contaminants that can further worsen your DNA sequence quality.

**Cause:** salts

**Solution:** excessive amounts of salts will also give rise to premature termination and may look similar to DNA overloading, with strong signal followed by progressively weakening signal. Salts have an inhibitory effect on the processivity of the sequencing Taq polymerase, which can lead to an overabundance of short fragments, or if the salt concentration is too great, the enzyme will be completely inhibited with no sequence data obtained. If salts are potentially a problem, perform an ethanol precipitation for salt removal.

**Cause:** repetitive regions



**Solution:** the nucleotide composition, as well as the size, of a repetitive region can play a large role in the success of sequencing through such an area. In general, G-C and G-T (often seen in bisulfite-treated DNA) repeats tend to be the most troublesome though, as mentioned before, the newest version of Applied Biosystems BigDye Terminator v3.1 contains some modifications that have allowed for some striking improvements in certain previously difficult templates. However, there are still some that remain a pain. In general, one can sequence partially through the repetitive region and the signal begins to fade and eventually becomes unreadable. This may be due to premature dNTP depletion, secondary structure formation or enzyme slippage. Various methods can be tried to sequence the repeat entirely, and many are similar to those we would use for G-C rich templates that form secondary structures. If the repeat region is not excessively large, sequencing from the opposite strand to complete the region can be successful, especially if the complementary strand has a nucleotide composition that is more efficiently extended. However, if the region is large, it may be difficult to complete its entire sequence and determine the exact number of repeats present. Alternative methods, such as directed deletions or the use of an in vitro transposon system may need to be utilized.

## Plasmid templates

### Considerations when cleaning up plasmid preps

Poor template quality is one of the most common reasons for bad sequence data, as mentioned above, and is a prime consideration when choosing a plasmid cleanup method to give DNA of optimal purity for automated sequencing. Plasmid template quality can be affected by a variety of factors and contaminants including the following:

- Salts or organics left over from template preparation
  - \* Presence of cellular components such as RNA, proteins, polysaccharides or chromosomal DNA
  - \* DNA that has degraded while in storage
  - \* silica fines that carryover from template preparation kits that utilize loose resin or silica solutions

## Contaminants

**Salts-** the processivity of the Taq polymerase used in the cycle sequencing reaction declines in the presence of high amounts of salts. Salt contamination in DNA preps may result from coprecipitation of salts in alcohol precipitations, insufficient removal of supernatant after precipitations or an incomplete wash of the pellet with 70% ethanol. Careful technique should be used when precipitating with alcohol. It has also been demonstrated that acetate ions, as opposed to sodium, potassium or chloride ions, are the most inhibitory in sequencing reactions. When using potassium acetate or sodium acetate, concentrations over 20 mM led to complete failure of the sequencing reactions, while concentrations of 60mM of sodium chloride were required before complete inhibition.

**Ethanol-** ethanol contamination can occur when the sample is insufficiently dried after precipitation or when carried over in an ethanol-containing wash buffer used in some DNA isolation procedures. Contamination with 10% or greater concentrations of ethanol usually leads to failure of the DNA sequencing reaction. Complete drying of the DNA samples is required to remove these traces of ethanol.

**Phenol-** phenol may be carried over from DNA alkaline lysis methods that utilize phenol and chloroform to remove proteins and other cellular contaminants from cell lysates. Phenol cannot

be tolerated in the cycle sequencing reaction as it denatures proteins and will thus degrade the Taq polymerase enzyme used in the cycle sequencing reaction. Chloroform does not have the strong denaturing properties of phenol and doesn't appear to adversely affect the sequencing reaction.

**EDTA** - EDTA can chelate the magnesium required by the Taq polymerase in the cycle sequencing reaction, so when submitting samples, it is best to always have them diluted or resuspended in sterile ddH<sub>2</sub>O or 1X Tris buffer. Suspension in TE buffer is not recommended, though people have done it and many times there is not a problem. However, providing template DNA in water is an easy thing to do and if there is a problem with your sequence quality, the fact that there is no EDTA in your sample is one potential problem we can eliminate right away.

**Finally, always check your chromatogram to be sure that the base caller did call true bases and not background peaks.**

This document was taken on the Roswell Park Institute website  
<http://www.roswellpark.org>